DOCUMENT RESUME

ED 385 558 TM 023 980

AUTHOR Kaplan, Randy M.; Bennett, Randy Elliot
TITLE Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypotheses Item. GRE Board Professional Report No. 90-02bP.
INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY Graduate Record Examinations Board, Princeton, N.J.
REPORT NO ETS-RR-94-08
PUB DATE Jun 94
NOTE 41p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Automation; *Computer Assisted Testing; Correlation; Higher Education; Hypothesis Testing; Responses; Scores; *Scoring; Semantics; *Test Items
IDENTIFIERS *Free Response Test Items; *Hypothesis Formulation; Pattern Matching

ABSTRACT

This study explores the potential for using a computer-based scoring procedure for the formulating-hypotheses (F-H) item. This item type presents a situation and asks the examinee to generate explanations for it. Each explanation is judged right or wrong, and the number of creditable explanations is summed to produce an item score. Scores were generated for 30 examinees' responses to each of 8 items by a semantic pattern-matching program and independently by 5 human raters. On its initial scoring run, the program agreed highly with the raters' mean item scores for some questions and improved its concurrence substantially as modifications to the automatic scoring process were made. By the final run, correlations between the program and the raters on item scores ranged from .89 to .97, and mean human-machine discrepancies ran from .6 to 1.1 on a 16-point scale. At the individual hypothesis level, the proportion agreement, given the large disproportion of correct responses in the sample, was little better than chance. F-H items might be more effectively scored by a semiautomatic system that combines machine processing with a small number of human judges, and a preliminary configuration for such a process is presented. Appendix A discusses scoring iterations and modifications to the tool, and Appendix B presents changes to the scoring tool's interface. (Contains 5 figures, 9 tables, and 14 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 385 558

GRE®
RESEARCH

Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypotheses Item

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Randy M. Kaplan
and
Randy Elliot Bennett

June 1994

GRE Board Professional Report No. 90-02bP
ETS Research Report 94-08

Tm023980

Educational Testing Service, Princeton, New Jersey

BEST COPY AVAILABLE

Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypotheses Item

Randy M. Kaplan
and
Randy Elliot Bennett

GRE Board Report No. 90-02bP

June 1994

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

****

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in Graduate Record Examinations Board Reports do not necessarily represent official Graduate Record Examinations Board position or policy.

****

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

EDUCATIONAL TESTING SERVICE, ETS, the ETS logo, GRADUATE RECORD EXAMINATIONS, and GRE are registered trademarks of Educational Testing Service.

Copyright © 1994 by Educational Testing Service. All rights reserved.

Abstract

Large-scale institutional testing, and testing in general, are in a period of rapid change. Among the more obvious dimensions is the growing use of constructed-response items and of computer-based testing. This study explores the potential for using a computer-based scoring procedure for the formulating-hypotheses item. This item type presents a situation and asks the examinee to generate explanations for it. Each explanation is judged right or wrong and the number of creditable explanations summed to produce an item score. Scores were generated for 30 examinees' responses to each of eight items by a semantic pattern-matching program and independently by five human raters. On its initial scoring run, the program agreed highly with the raters' mean item scores for some questions and improved its concurrence substantially as modifications to the automatic scoring process were made. By the final run, correlations between the program and the raters on item scores ranged from .89 to .97, and mean human-machine discrepancies ran from .6 to 1.1 on a 16-point scale. At the individual-hypothesis level, the proportion agreement ranged from .80 to .94, which, given the large disproportion of correct responses in the sample, was little better than chance. Also detected was a tendency on the part of the program to erroneously classify wrong responses as correct. We conclude that F-H items might be more effectively scored by a semiautomatic system that combines machine processing with a small number of human judges, and we present a preliminary configuration for such a process.

Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypotheses Item

Large-scale institutional testing, and testing in general, are in a period of rapid change. Among the more obvious dimensions of this change is the growing use of constructed-response items and of computer-based testing. Considerable research has been carried out on how to grade responses to open-ended items automatically. If accurate methods of semiautomatic or automatic scoring can be devised, large-volume computer-based programs can begin to use tasks that more faithfully replicate the kinds of problems examinees face in academic and work settings.

Some success in creating programs for automatic analysis has been achieved. Architectural site designs, brief computer programs, the equations leading to the solution of algebra word problems, and sentence-length answers to reading comprehension questions have each been graded with accuracy approaching that of human content experts (Bejar 1991; Braun, Bennett, Frye, & Soloway, 1990; Kaplan, 1992; Sebrechts, Bennett, & Rock, 1991).

This report describes an attempt to score responses to formulating-hypotheses (F-H) items automatically. The F-H task was created by Frederiksen (1959) to measure the ability to interpret research results as a scholar would. The item type presents a situation and asks the examinee to generate explanations for it. Early studies showed the task to measure a divergent thinking ability different from the constructs tapped by the GRE General Test and to predict some types of accomplishment better than that test (Frederiksen & Ward, 1978; Ward, Frederiksen, & Carlson, 1980). Attempts to make a machine-scorable multiple-choice version failed because the resulting score correlated too highly with the General Test (Ward, Carlson, & Woisetschlaeger, 1983). Interest in the item type was renewed for two reasons. First, the ability to generate alternatives was found to be important to success in graduate education (Powers & Enright, 1987; Tucker, 1985). Second, the progress in computer-based natural language processing suggested that automatic scoring might be feasible using a semantically based pattern-recognition approach (Carlson & Ward, 1988).

An important step toward an automatically scorable version of F-H was realized in the development of the Semantic Pattern Matching Scoring Program (**SPAM-SCOR**) (Kaplan, 1992), which was capable of grading sentence-length natural-language responses. **SPAM-SCOR** was applied to free responses of reading comprehension items. It was able to duplicate the judgments of a human grader perfectly for two items with average response lengths of 3-5 words and agreed with graders on 88% of the answers to an item whose average response was 12 words long.

SPAM-SCOR's success set the stage for building a computer-based version of F-H and exploring the feasibility of automatically grading the results. Bennett and Rock (1993) reported on the validity of scores from that computer-based version, in general replicating the findings from the earlier paper-and-pencil F-H studies. In the current report, we briefly describe the computer-based F-H test, introduce a program for automatically scoring responses, evaluate the accuracy of that automatic analysis, and give a general outline for a production scoring process and the research needed to develop it.

The Computer-Based Formulating-Hypotheses (F-H) Test

The computer-based formulating-hypotheses prototype was developed to determine if the item type could be effectively computer delivered and machine scored. The F-H test consisted of eight items. In general, these items required no specific disciplinary knowledge but, rather, general knowledge about the world. Responses to each item were constrained to either a maximum of 7 or a maximum of 15 words. Among other things, this limitation was imposed to explore the effect on scoring accuracy, which, in earlier work, was higher for shorter responses (Kaplan, 1992).

The interface used to present F-H questions and collect examinee responses is illustrated in Figure 1. The top left-hand window shows an item, with directions for completing the task given in the bottom left window. The examinee types a hypothesis, which appears in the lower right box. When the SAVE button is clicked with the mouse, the hypothesis is moved to the list in the upper right-hand window. To edit a hypothesis on the list, the examinee highlights it with the mouse and clicks on the EDIT button, moving the hypothesis back to the entry box where it can be changed.

Each F-H item is scored on a 0-15 scale with one point awarded for each plausible, unduplicated hypothesis, as defined by a scoring rubric developed as part of a related validity study (Bennett & Rock, 1993). The rubric lists several general categories--and within these, several specific categories--into which correct responses might fall. In general, a response is considered creditable if it states or implies a possible explanation that is readily apparent and does not duplicate another hypothesis generated by the student for that problem. Duplication is defined as more than one hypothesis falling into the same

specific category. In addition to duplication, a response is not to be credited if it directly contradicts the situation or if it is clearly implausible.

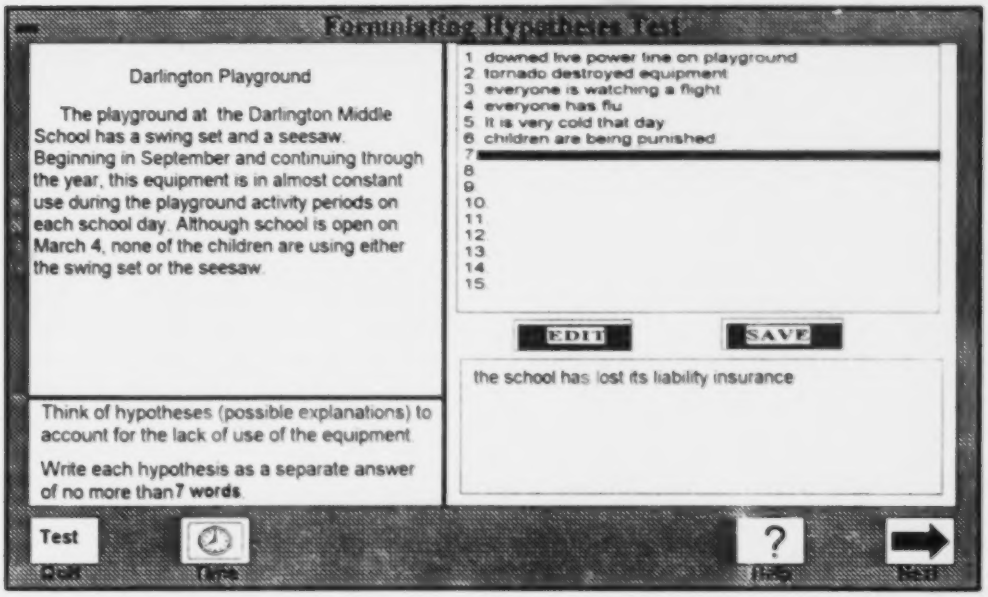
Formulating Hypotheses Test
Darlington Playground
The playground at the Darlington Middle School has a swing set and a seesaw. Beginning in September and continuing through the year, this equipment is in almost constant use during the playground activity periods on each school day. Although school is open on March 4, none of the children are using either the swing set or the seesaw.
Think of hypotheses (possible explanations) to account for the lack of use of the equipment.
Write each hypothesis as a separate answer of no more than 7 words.
1 downed live power line on playground
2 tornado destroyed equipment
3 everyone is watching a flight
4 everyone has flu
5 it is very cold that day
6 children are being punished
7
8
9
10
11
12
13
14
15
EDIT
SAVE
the school has lost its liability insurance
Test
?

Figure 1 - F-H interface screen

The Free Response Scoring Tool (FRST)

FRST (Free-Response Scoring Tool) is a pattern-based program for scoring length-limited or domain-limited natural language responses. **FRST** is the evolution of **SPAM-SCOR** (Kaplan, 1992). Both programs are trained on a sample of responses to induce a grammar describing the larger set of responses. The induced grammar is then used as the scoring key. If the responses of the training set represent correct responses, and **FRST** recognizes a response using the induced grammar, that response is part of the training set and therefore part of the set of creditable hypotheses (excepting duplicates, which **FRST** cannot currently detect).

The **FRST** scoring process involves several fundamental steps, repeated for each new item (see Figure 2). (A complete description of this process is contained in Kaplan, 1992.)

BEST COPY AVAILABLE

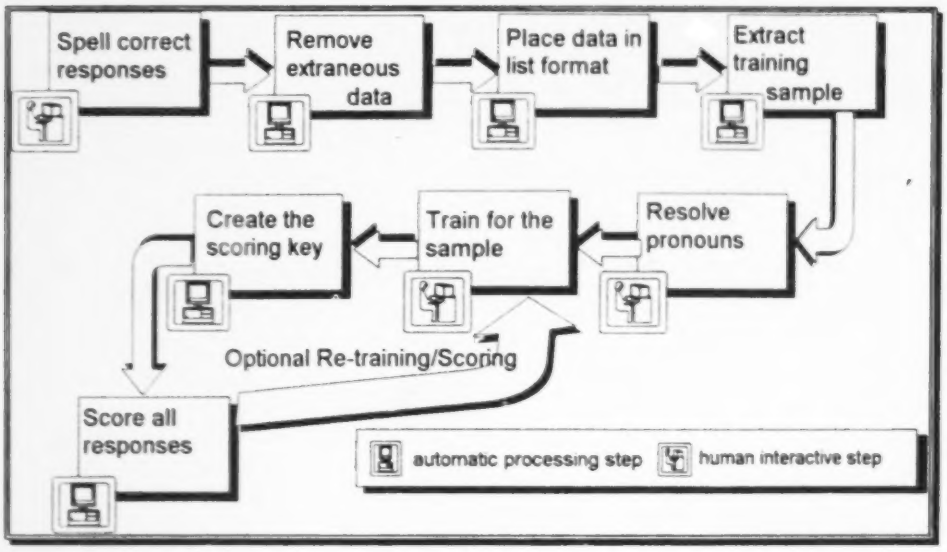

Figure 2 - **FRST** processing

In the first step, manual spell correction of responses is done to permit words or phrases to be properly classified. Punctuation is then removed, numbers transformed into words, and the data placed into list format.

Next, responses are selected for training **FRST** to score the item. The training sample must be representative of the range of structure and vocabulary that might occur in responses for that item.

Following this, pronouns are manually resolved. For example, in an item dealing with improvements in job safety for police officers, this step would involve translating "They wore bulletproof vests" into "Police wore bulletproof vests."

Training follows pronoun resolution. In **FRST** nomenclature this step is called stage one. During stage-one, a test developer interacts with the program to create a scoring key. The process consists of examining each response, selecting one or more elements, and assigning semantic classes to those elements. An element may be a word, a phrase, or the complete response. The last is for the case where the response forms a semantic class by itself. For example, consider the Police Officers item and the response that follows it:

The job of a police officer in the 1990's appears to be much less hazardous than it was 20 years ago. In 1970 more than 3.5 police per thousand officers were killed in the line of duty, but by 1989 the rate dropped to only 1.5 deaths per thousand.

Police wore bulletproof vests.

From this response, two phrases will be assigned to semantic classes:

police
bulletproof vests

Police will be assigned to the class **law enforcement personnel** and **bulletproof vests** will be assigned to the semantic class **protective clothing**. The semantic classes are created during the training process as the need for them arises.

Once training is complete, the scoring key is created. This step is automatic and does not require any user intervention. In this part of the process, the training responses are canonicalized according to the semantic classes. A response canonicalization consists of modifying an original response to remove any elements that were specified as not relevant during the training process and to replace elements with others in the lexicon. The canonicalized responses are then converted into patterns, and these patterns are organized into **FRST's** scoring key (the induced grammar). In **FRST** this is called **stage two**.

Samples of original responses, canonicalized responses, and the scoring key patterns for the Police Officers item are shown in Table 1.

Table 1
Sample of raw and canonicalized responses

Raw Response	Canonicalized Response	Raw Scoring Key Pattern
better laws protecting policemen	better laws protecting policemen	police protect better
policemen are better trained	policemen better trained	better police education
medical science is better able to save these officers	medical better save officers	medical better save police
better bulletproof vests have been developed	better bulletproof vests	bulletproof vests
less strict laws for policemen who shoot first	less laws policemen who shoot first	police shoot first

From Table 1, one can see that a canonicalized response is the basis for a scoring key pattern. Specifically, once a response is canonicalized, the words or phrases in that response are replaced with their corresponding semantic class names. For example, from Table 1, the response

policemen are better trained

is canonicalized to

policemen better trained

and the raw scoring key pattern corresponding to this canonicalized response is

better police education

Although the specific words used in the canonicalized response and the key pattern are sometimes the same, they are interpreted differently by **FRST**. In the case of the canonicalized response, the words still represent words and phrases of the original response. In the case of the scoring key, these words are the names of semantic classes. **police** in this pattern is the name of the semantic class that contains all words or phrases that could represent a law enforcement official. Likewise, the term **better** refers to any word or phrase that connotes improvement. The transformation from response to canonicalized response to a raw scoring pattern is done for each response in the training set. The key is composed of all distinct raw scoring patterns (duplicates are removed), ordered by length of the pattern.

In the event a word or phrase is classified into more than a single semantic class, the context of the classification is stored with the classification. The context consists of the responses of the training set for which the classification was made.

If a word or phrase has no semantic classification associated with it, an attempt is made to assign one using a dictionary that is available to **FRST**. If, after the dictionary has been used, no semantic classification can be found, the word will be removed from the response before scoring.

The scoring key created by **FRST** is related to the rubrics created previously by test developers (Bennett & Rock, 1993), but it has some important differences. An excerpt from the test developers' rubric for the Police Officers item is shown in Figure 3.

GENERAL CATEGORY	SPECIFIC CATEGORY
A) BETTER PROTECTION	1. bulletproof vests 2. brighter clothes for traffic police 3. police dogs
B) BETTER WEAPONS / EQUIPMENT	1. guns that shoot faster, easier to aim 2. smoke bombs, mace, etc. 3. police cars, etc., safer
C) BETTER PROCEDURES / TECHNIQUES	1. work with partner / in larger groups 2. stay in cars, not on street beat 3. call for more help 4. use bullhorns 5. better intervention / counseling for crooks 6. trained to shoot faster 7. trained for self-defense / safety
D) MEDICAL ADVANCES	1. more wounded police saved
E) BETTER POLICE	1. smarter, higher IQ's, more mature 2. more honest, less connected to criminals 3. more fit, athletic, healthier, alert

Figure 3 - Excerpt from test developer rubric for Police item

The test developer rubric consists of a series of general categories into which responses are classified. Each of these categories may have one or more specific categories. As can be seen in Figure 4, the key used by **FRST** is more specific than the rubric created by the test developers. **FRST** must be trained for the features that lead a human rater to classify a response into a particular category. These features are the basis for the relationship between the key used by **FRST** and the human raters' rubric.

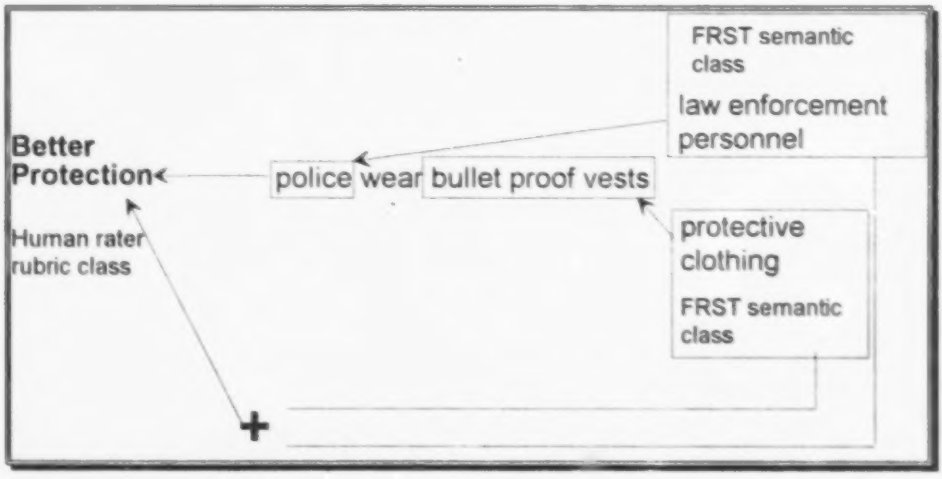

Figure 4 - Relationship of human rater rubrics to **FRST** rubrics

The test developers' rubric supplies the general response categories and examples of specific categories for training **FRST**. Both the general and specific categories supply vocabulary for the training process. For example, the category **better protection** supplies the following vocabulary:

bulletproof vests
brighter clothes
police
dogs
better protection

This initial vocabulary is assigned to one or more semantic classes. This is shown in Figure 4. **Police** is assigned to the semantic class **law enforcement personnel**, and **bulletproof vests** is assigned to the semantic class **protective-clothing**. When both of these elements appear in a response, the response is classified as part of the training set because, together, these elements support the membership of the response in the general category **better protection**.

We say initial vocabulary because it is possible there will be other responses that also are part of the general category **better protection** but that do not use exactly the same vocabulary. For example, consider the response

Police wear more keflar clothing

The word **keflar** is not part of the human raters' rubric, but we know from common knowledge that keflar is used to make bulletproof vests. Encountering this response during training would result in the classification of keflar in the semantic class **protective clothing**. Just as a test developer uses the rubric as a guide for scoring an F-H response, the **FRST** trainer will use the rubric to classify response elements.

After stage-two training is complete and the key file has been created, the last step is to score the responses. Scoring consists of canonicalizing the response and creating a pattern, and looking to see if the pattern is part of the scoring key. If it is part of the scoring key, then the response corresponding to that pattern is recognized[1] as part of the training set.

One important heuristic used during the scoring process is the order relaxation heuristic. This heuristic permits a higher level of recognition by **FRST**. The heuristic

[1] In this study, recognized refers to those responses that are considered to be correct for the item. These are the only responses that were trained for.

allows a match to take place between a pattern in the scoring key and the pattern of a response, even if the order of semantic classes in the scoring key pattern does not match the order of semantic classes in the response pattern. For example, consider the following two responses from the Police Officers item:

better trained officers
police training in safety precautions has improved

The canonicalized versions of these responses are

better trained officers
police training improved

Both of these responses would be classified by the same pattern in the scoring key, namely,

better trained officers

The first response canonicalization matches this pattern exactly. The second matches if the elements of the response are rearranged

improved training police

The ordering heuristic reorders responses in just this way and will allow any ordering to be accepted.[2]

A second heuristic is called the additional information heuristic. If a response pattern matches a pattern in the scoring key, and if the response pattern also has additional semantic classes (not in the scoring key pattern), the patterns are regarded as matching and the additional semantic classes are ignored. For example, consider the following response from the Police Officers item:

More effective means of training has reduced death rate of officers on duty

and its corresponding canonicalization:

---

[2] Even though these responses are canonicalized into the same pattern, semantically they are different. The issue of duplicates is an important one for scoring responses to the F-H item. Although FRST currently has no mechanism for detecting duplicates, one basis for identification exists in the canonicalization and pattern-matching process (as indicated in this example). That is, different responses may reduce to the same response pattern or may match the same key pattern through the use of heuristics. An important issue is the extent to which human judges would agree that such hypotheses are, in fact, redundant and, if not, how to build protections into the scoring process to prevent semantically different hypotheses from being treated as equivalent.

more effective training reduced death officers

Scoring would transform this response into the following semantic pattern:

better training less death police

The scoring key contains the pattern

less death

Because this matches a subpattern of the response pattern, the response is recognized as one described by the training set. This heuristic is used only when there is no scoring key pattern whose length is the same as a response pattern and whose semantic classes exactly match those of the response. In general, both the order relaxation and additional information heuristic were found to be effective in the earlier study (Kaplan, 1992).

Evaluation

Method

Subjects. Subjects were participants in a study to evaluate the validity of a computer-based F-H test (Bennett & Rock, 1993). For that study, data were gathered from 211 paid volunteers recruited by contacting graduate departments at institutions near 12 Educational Testing Service (ETS) computer-based test centers in the Southern, Western, Midwest, Mid-Atlantic, and Northeast regions of the United States. Contacts were made primarily through education, psychology, English, chemistry, and biology departments. Included in the participation guidelines were the following qualifications: students had taken the GRE General Test during the 1990-1991 academic year, were already enrolled in the first year of a graduate degree program, and were native English speakers. The current study used a subsample of 30 examinees drawn from this data set for its primary analyses.

Table 2 shows how this graduate sample compared demographically with the 1987-88 GRE General Test examinee population, the most recent one for which data were available. As might be expected, the sample diverged from the test population in noticeable ways. The sample scored considerably higher on the three General Test scales and had proportionally more females, U.S. citizens, individuals whose graduate objective was the Ph.D., and social science and humanities/arts majors. Physical science and engineering majors and majors classified as "other" were underrepresented.

Table 2
Demographic Data

Background Characteristic	Study Sample (n=30)[a]	1987-88 Examinee Population (n > 185,000)
General Test Performance		
Verbal mean (SD)	587 (123)	486 (122)
Quantitative mean (SD)	609 (127)	553 (139)
Analytical mean (SD)	614 (138)	529 (128)
Percentage Female	62%	53%
Percentage Non-White	15%	14%
Percentage U.S. Citizen	90%	81%
Percentage with Ph.D. Goal	69%	40%
Graduate Major		
Social Sciences	48%	18%
Humanities/Arts	17%	11%
Life Sciences	17%	18%
Education	13%	15%
Physical Sciences	4%	11%
Engineering	0%	12%
Business	0%	3%
Other	0%	12%

Note. Population data are from Examinee and Score Trends for the GRE General Test by D. M. Wah and D. S. Robinson, Copyright 1990 by Educational Testing Service. Percentage Non-White is for U.S. citizens only. Graduate major percentages for population are based on those with decided majors only.

[a]The percentages for Non-White, U.S. citizen, Ph.D. objective, and graduate major are based on n's of 26, 29, 29, and 23 respectively.

Procedure and data analysis. FRST's accuracy was evaluated by examining its agreement with human raters. As part of the Bennett and Rock (1993) study, the responses of the above-described 30 examinees were given in hard copy form to four ETS test developers and one ETS consultant to score. Four of the readers had earned an M.A. and one a Ph.D.; three had majored in English literature, one in education, and one in physics. Before each item was scored, the rubric was introduced, sample responses were discussed, and several responses were graded for practice purposes. All five readers then independently graded all 30 responses. This process was repeated until all eight items had been evaluated. A variance components analysis was then computed. The results suggested that judges generally agreed on the scores they produced for F-H items. The overwhelming majority of the total variance was attributable either to persons (72% for 7-word items, 63% for 15-word items), or to the persons x question interaction. (This latter component, accounting for 22% of the total variance for the 7-word items and 24% for the 15-word ones, indicated that some examinees did well on some items but poorly on others.) Generalizability coefficients for the mean ratings taken across judges and items were .93 for the F-H 7-word items and .90 for the 15-word items, indicating that the raters provided a stable standard against which to compare **FRST**.

As reported in Bennett and Rock (1993), the raters credited most hypotheses that examinees offered (see Table 3). That the preponderance of responses should be considered correct is reasonable, given that F-H items have many right answers by design. However, as a consequence, this data set offers only limited opportunity to test **FRST**'s ability to detect wrong responses accurately.

Table 3
Number of Credited versus Number of Offered Hypotheses (n=30)

Item	Mean # of Hypotheses Offered	Mean # Credited by Raters
F-H 7-Word		
1. Darlington Playground	9.7	8.8
2. Mackerel Catch	9.7	9.2
3. Missing Daybooks	9.1	8.5
4. Datar's Beaches	9.0	8.4
Total	37.5	34.8
F-H 15-Word		
5. Minor Dutch Painters	7.6	6.5
6. Kingston Deer	8.8	7.9
7. Police Officers	8.9	8.1
8. Disease in Alcadia	8.8	8.2
Total	34.2	30.7

Note. The mean of the number of hypotheses credited is taken across examinees and raters.

For the present investigation, **FRST** was trained to score this study sample using an independent group of 45 examinees. 23 examinees used in the Bennett and Rock (1993) validity study, 9 cases excluded from that study for such reasons as loss of auxiliary data, and 13 cases from a local pilot test. (This training sample was also used to develop the scoring rubrics employed by the human judges in the Bennett and Rock investigation.)

After training, **FRST** was run against the study sample, the correlation of its scores with the raters was checked, and, if the correlation was substantially below .90, one or two additional runs were made. (In general, **FRST**'s performance on the individual responses composing the study sample was not examined.) Before making additional runs **FRST** was either modified (e.g., by adding a morphological analyzer) or retrained using techniques that had worked more successfully for other items (see Appendix A for a description of the changes). Because the results of these additional runs were not cross-validated, we report both initial and final outcomes, with the expectation that **FRST's** performance on a new examinee sample would probably lie somewhere in between.

Agreement between **FRST** and the raters was computed in four ways. First, the program's mean scores were compared with the rater means to identify any systematic

leniency or strictness. Next, the Pearson product-moment correlations between **FRST**'s scores and the mean scores taken across raters were calculated. This mean score is conceptually similar to classical test theory's "true" score, the mean of many independent observations of the same performance and, as such, is an approximation of what the "correct" score for an examinee should be. Third, the discrepancies between **FRST**'s scores and the rater means for each examinee were evaluated. Finally, **FRST**'s right/wrong categorizations of individual hypotheses were compared to the human judgments. For this last analysis, duplicates were counted as correct, as **FRST** currently has no mechanism for detecting redundant hypotheses and because such responses are, by definition, repetitions of other correct responses.

Results

Table 4 gives summary statistics for **FRST**'s and the human raters' item and total scores. For **FRST**, values for initial and final runs are displayed for the applicable items. Given for the human raters are the mean scores taken across raters and examinees and the means of the score standard deviations (where each standard deviation was taken across examinees for a rater). For the initial run, the machine and human distributions diverged considerably for five of the eight items, with the machine scores generally being lower and more narrowly distributed. The two distributions were more similar for the final run, for which retraining was done or **FRST** was changed (see Appendix A).

Table 5 presents the Pearson product-moment correlations among the rater mean scores, **FRST**'s scores, and the number of hypotheses offered.[3] For the initial run, **FRST's** agreement with the human judges varied widely from .33 to .97. Values for the final run were consistently high, ranging from .89 to .97 for the item scores, .98 for the 7-word total score, and .96 for the 15-word total score. As noted, the reasons for the improvement were the changes to **FRST**.

As the table also shows, the rater scores were predicted as well or better by a simple count of the number of hypotheses offered. This relationship again suggests that the main function of a sophisticated scoring program for F-H will be not so much to score the typical examinee's productions (for which a simple count would be almost as good as a more careful analysis), but to discourage offering extra (but erroneous) responses in an attempt to improve one's scores.

[3] Note that the rater scores discount duplicates, whereas FRST's scores and the number of hypotheses offered do not.

Table 4
Summary Statistics for **FRST** and Human Rater Item Scores (n=30)

Item	Machine Scores				Human Rater Scores	
	Mean		SD			
	Initial Run	Last Run	Initial Run	Last Run	Mean	Mean SD
F-H 7-word						
1. Darlington Playground[a]	5.6	8.2	2.0	3.4	8.8	3.4
2. Mackerel Catch[b]	3.6	8.3	2.2	3.3	9.2	3.4
3. Missing Daybooks	8.6	---	3.4	---	8.5	3.4
4. Datar's Beaches	8.0	---	3.4	---	8.4	3.5
Total	25.7	33.1	9.1	12.1	34.8	12.3
F-H 15-word						
5. Minor Dutch Painters	5.8	---	3.1	---	6.5	3.1
6. Kingston Deer[a]	5.6	7.8	2.0	3.1	7.9	3.2
7. Police Officers[a]	6.9	8.3	3.2	3.1	8.1	3.2
8. Disease in Alcadia[a]	5.5	7.8	2.7	3.4	8.2	3.6
Total	23.9	29.7	8.4	10.4	30.7	11.4

Note. The mean of the number of hypotheses credited is taken across examinees and raters. The mean SD is the average of the examinee score standard deviations, one for each rater.

[a]Values are for each of two runs.
[b]Value are for first and last of three runs.

Table 5
Correlations Among **FRST** Scores, Human Rater Scores, and Number of Responses Offered (n = 30)

Item	**FRST** with Rater Mean		# Offered with Rater Mean	**FRST** with # Offered	
	Initial Run	Last Run		Initial Run	Last Run
7-word					
1. Darlington Playground[a]	.65	.95	.97	.67	.94
2. Mackerel Catch[b]	.33	.92	.99	.35	.90
3. Missing Daybooks	.97	---	.99	.97	---
4. Datar's Beaches	.95	---	.99	.96	---
Total	.91	.98	.99	.92	.98
15-word					
5. Minor Dutch Painters	.93	---	.94	.86	---
6. Kingston Deer[a]	.72	.89	.96	.74	.91
7. Police Officers[a]	.78	.92	.96	.82	.96
8. Disease in Alcadia[a]	.66	.96	.99	.64	.95
Total	.85	.97	.97	.83	.95

[a]Values are for each of two runs.
[b]Values are for first and last of three runs.

Table 6 summarizes the discrepancies between **FRST**'s scores and the mean of the raters' scores, where a discrepancy was calculated by subtracting **FRST**'s score from the rater mean. The mean absolute difference indicates how far off **FRST**'s scores were on average, and the mean signed difference gives the direction and magnitude of any systematic bias. As expected from the distributional and correlational results, discrepancies varied widely for the initial run, from absolute values for items of well under 1 point to over 5 points (on a 16-point scale). The final-run values were more acceptable: For items, the mean absolute differences ran from .6 to 1.1, or from 4% to 7% of the 16-point score scale. Mean signed differences ranged from -.9 to .9. Finally, note that for the last run there was little difference in scoring accuracy for the 7-word versus 15-word items. The mean absolute discrepancies for both total scores represented 4% of the 60-point scale.

Table 6
Discrepancies Between **FRST** Scores and Human Rater Mean Scores

Item	Mean Absolute Discrepancy		Mean Signed Discrepancy	
	Initial Run	Last Run	Initial Run	Last Run
7-word				
1. Darlington Playground[a]	3.1	.9	3.1	.5
2. Mackerel Catch[b]	5.6	1.1	5.6	.9
3. Missing Daybooks	.6	---	-.1	---
4. Datar's Beaches		.9	.4	---
Total	9.1	2.4	9.1	1.7
15-word				
5. Minor Dutch Painters	.9	---	.7	
6. Kingston Deer[a]	2.4	.9	2.3	.1
7. Police Officers[a]	1.6	.9	1.2	-.2
8. Disease in Alcadia[a]	2.7	.7	2.6	.4
Total	6.8	2.4	6.8	1.0

Note. Positive differences indicate that the judges' mean score was higher than **FRST**'s score.

[a]Values are for each of two runs.
[b]Values are for first and last of three runs.

Although the final run reduced most of the individual discrepancies to minimal levels, there were several quite substantial deviations. Nine of the 240 deviations in item scores exceeded three points. Differences of this size occurred for all items except 3 and 8. The discrepancies whose deviations were more than three points are summarized in Table 7.

Table 7
Summary of Discrepancies by Items

Item	Subject Id	Discrepancy	Missing Pattern	Error in Canonical-ization	Duplicate response undetected
Darlington Playground	274	3.4	2	2	
Mackerel Catch	219	5.8	3	3	
Mackerel Catch	266	3.0	4		
Datar's Beaches	192	3.4	5	1	
Minor Dutch Landscape Painters	192	3.8	4	1	
Minor Dutch Landscape Painters	093	3.0	3	1	
Kingston Deer	114	5.4	6		
Police Officers	266	-4.0			4
Police Officers	1001	-3.0			3
Totals			27	8	7

Three problems primarily account for the discrepancies listed in Table 7: **missing patterns, canonicalization errors, and undetected duplicate responses**. By far, the most frequent cause of the discrepancies is missing patterns.

Missing patterns are caused when the training set does not account for some word or phrase that appears in the responses for an item and this cannot be resolved by morphological analysis or dictionary search. The remedy for this problem would be to increase the training sample to include a pattern for the particular missing word or phrase. (In a production setting, this problem could be remedied by adding the pattern to the scoring key when it occurs during the scoring process, a notion we discuss later.)

The second problem, canonicalization errors, appears when a response has no canonicalization even when it should have, or when the canonicalization fails to capture essential elements of a response. A response will fail to have a canonicalization when no word or phrase in that response has a semantic class associated with it. This problem is distinguished from the missing-pattern problem in that, if the canonicalization process used dictionary lookup and morphological analysis, the response would very likely have a recognizable pattern associated with it. In the present version of **FRST**, the canonicalization procedure uses no dictionary search or morphological analysis. An important modification would be to add these two processing elements during stage two and score processing. For example, consider the following hypothesis from the Mackerel Catch item. In the missing-pattern case, the response never occurred and therefore no training was carried out for it.

The fleets went sailing somewhere else.

No canonicalization was produced for this response. It may have been that the lexicon contained, for example, **ships** as an entry. Because the dictionary is not called upon at canonicalization time, no check to see if **fleets** is a synonym for **ships** would be done. No lexicon entries were made for the rest of the response. Because the result of this is no canonicalization, this response will not be recognized even though it might have been correctly processed with morphological analysis or dictionary lookup.

In the case where **FRST** assigned a score greater than that of the human raters, it was frequently because of undetected duplicate responses. For example, for the Police Officers item, the following hypotheses were given in the same response and rated by at least one human rater as duplicate:

Police have more backup to shoot first.
Police have shot before being shot.
Less strict laws for policemen who shoot first.

When these were scored in **FRST**, they were all recognized, and all contributed to the total score. **FRST** cannot discern that they are versions of the same hypothesis. The ability to detect duplicates could be added to **FRST** by using the semantic class data obtained during training.

Table 8
Summary of #114, #192 and #266 responses

Subject #	Item	Discrepancy > 0	Hypotheses not scored by **FRST**, but recognized by at least 4 of 5 human raters	**FRST** Problem
114	Kingston Deer	5.4	there was exceptionally good road/weather conditions in 1986	np
			there was a drop in tourism through Kingston in 1986 due to recession	np
			there was a huge forest fire that year (4/5)	np
			gasoline prices soared that year	np
			the roads were flooded out that year	np
			there was a position turnover for the record keeper that year	np
192	Datar's Beaches	3.4	increase of fishing industry	np
			dumping increase (4/5)	np
			improved means of reporting information	np
			deterioration of methods of reporting information(4/5)	np
			proliferation of paper	ce
192	Minor Dutch Landscape Painters	3.8	fashionable to attribute authorship to major artists	np
			little is known about the minors-proof difficult	ce
			conflicting bills of sale	np
			poor scholarship(4/5)	np
			attempt to build reputation of conservators-no one wants to work with minors(4/5)	np
266	Mackerel Catch	3.0	the mackerel became dangerous to eat	np
			the port suddenly became a dangerous area	np
			someone else took over the port administration	np
			most fisherman went off to the army	np
			a rash of robberies occurred	np
			another type of fish became more desirable	np
Subject #	Item	Discrepancy < 0	Hypotheses scored by **FRST**, and scored as duplicate by at least 1 human rater	**FRST** Problem
266	Police Officers	-4.0	policemen are better trained	
			policemen are better skilled at those situations	ud
			policemen have better developed skills at those situations	ud
			stiffer penalties for shooting an officer	
			the death penalty has been enforced	ud
			better bullet proof vests have been developed	
			policemen wear more protective garment	ud
			police have more backup to shoot first	
			less strict laws for policemen who shoot first	ud

Note: np = no pattern, ce = canonicalization error, ud = undetected duplicate.

Two subjects accounted for a plurality of discrepancies (see Table 8). When the discrepancy was positive, either the pattern was missing from the scoring key or there was a canonicalization error, suggesting that these examinees tended more than others to pose either substantively or linguistically unusual hypotheses that **FRST** did not recognize. These have been marked in the table. In the case of a negative discrepancy, where **FRST** considered more of the responses correct than did human raters, this was partly because **FRST** could not detect duplicates.

Because the mean discrepancy indices reported above miss cases in which disagreements over the scoring of individual hypotheses cancel out, we also analyzed **FRST**'s decisions for individual hypotheses. Each hypothesis was classified as right or wrong based on the judgment of the majority of raters (i.e., three or more). (Duplicates were coded as correct for this analysis because FRST has no mechanism for detecting them and, as individual responses, they are conceptually correct.) Each of the judges' right/wrong classifications was then compared to **FRST**'s judgment for that response.

Several indices were used to summarize the results. The proportion correct is the number of agreements between **FRST** and the raters divided by the number of agreements and disagreements. The false positive rate is the percentage of total hypotheses **FRST** erroneously considers to be right, and the false negative rate is the percentage **FRST** erroneously considers to be wrong. Both rates are affected by the split of true positive and true negative responses. The number of hypotheses correctly designated as right by **FRST** divided by the number considered right by the humans is the sensitivity. Specificity is the number correctly classified as wrong by **FRST** divided by the number discredited by the humans. Specificity and sensitivity are unaffected by the split of true positives and true negatives. Finally, kappa is a measure of agreement beyond that expected by chance. Normally, significant kappa values greater than .75 may be taken to represent excellent agreement beyond chance, values between .40 and .75 represent fair to good agreement, and values below .40 represent poor agreement beyond chance (Landis & Koch, 1977). However, in the current case multiple hypotheses were generated by the same individuals, so the rated responses are not necessarily independent. Thus, kappa's standard errors may be imprecisely estimated, suggesting that kappa be taken as a rough guide rather than an absolute indicator.

Table 9 gives results. **FRST** was able to correctly classify most of the responses it encountered, as indicated by proportions correct of between .80 and .94. (Note that because the judgment of the majority of raters was used to categorize each response, the raters necessarily agreed among themselves in 100% of instances.) The associated kappa values are, however, uniformly low, indicating little if any agreement beyond chance. The false positive rates are also invariably low, whereas the false negative rates are generally higher. This result occurs because of the disproportionate split of true positives and true negatives, which provides many more opportunities for making false negative than false positive errors. The sensitivity and specificity indices also diverge from one another, but more dramatically. The former closely track the proportion correct (because the number

of true rights is almost equal to the total number of responses), whereas the latter are usually much lower. This divergence needs to be interpreted cautiously because the specificity indices are very poorly estimated, given the extremely small numbers of true wrong responses for some items. The trend, however, suggests that **FRST** was better able to identify true right responses than true wrong ones (though again the large disproportion of right responses caused it to misclassify rights in greater numbers than it misclassified true wrongs). For item 7, for example, the program correctly identified 93% of the right responses (235 of 252) but only 6% of the wrong ones (1 of 16). Last, note that the two total scores produced values generally similar to one another.

Table 9
Accuracy of Right/Wrong Scoring Decisions for Individual Hypotheses

Total Score	n	Proportion Correct	False Positive Rate	False Negative Rate	Sensitivity	Specificity	Kappa
7-word							
1	291	.85	.00	.15	.85	.50	.03
2	292	.87	.00	.13	.87	1.00	.12
3	273	.94	.01	.05	.95	.50	.25
4	270	.89	.00	.11	.89	.50	.05
Total	1126	.88	.00	.11	.89	.62	.09
15-word							
5	228	.80	.01	.18	.80	.79	.26*
6	265	.88	.02	.10	.90	.50	.22
7	268	.88	.06	.06	.93	.06	.00
8	265	.86	.03	.11	.89	.25	.08
Total	1026	.86	.03	.11	.88	.39	.16*

* $p < .05$

This analysis of individual hypotheses gives a somewhat different picture of **FRST**'s accuracy than the analysis at the item-score level. This is partly because item scores are aggregations that allow errors in grading individual hypotheses to cancel out. It is also because the data set contains so few wrong responses. This low base-rate situation permits **FRST** to capitalize on chance: Overly general pattern matching (as indicated by low specificity) helps **FRST** recognize most responses as correct, producing high agreement with human judges because most responses are, in fact, correct.

Discussion

This study evaluated **FRST's** potential to automatically score responses to the F-H item. The F-H item offers a unique opportunity to test **FRST's** utility because F-H requires recognizing that responses belong to a specified set, the task **FRST** was primarily designed to perform.

We have shown that (1) **FRST** can score responses to these items, (2) its item-level scores can agree with human judgments reasonably well, and (3) the discrepancies can be explained in ways that, we believe, can be addressed. These findings suggest that **FRST** might be used as part of a semiautomatic scoring system.

A semiautomatic scoring system for F-H would consist of automatic and manual components. What would we want in the automated component of this system? Because the majority of F-H responses posed by examinees are correct, we would want this component to identify all true wrongs correctly (i.e., classify no wrong answers as right). Under this condition, responses scored as right would require no human verification (with the exception of sampling for quality control). However, because some right responses would be erroneously classified as wrong, only responses scored as incorrect would need to be verified. Because the number of true wrongs is very small to begin with, the extent of human involvement would, it is hoped, be very limited.

With the current data set, **FRST** accurately detected a large proportion of true correct responses but detected only a small proportion of the true wrong ones. This result can be attributed to training that made the program's keys too general. Because its matching was too general, it erroneously considered most wrong responses to be right, producing high sensitivity but low specificity.

In general, **FRST** needs to be more stringent in its pattern matching. This can be achieved by training to produce more specific key patterns, eliminating the special generalized pattern-matching capability used in this study, and manually updating its keys in the course of scoring (which we discuss below). Our results might also be improved through retraining and respecifying the keys in concert with test developers so as to use their scoring criteria in the training process, having them verify the final keys before scoring, integrating dictionary search and morphological analysis more fully, and using single-word patterns and general heuristics more cautiously. Finally, **FRST** training could incorporate both correct responses and incorrect responses (as opposed to only correct ones). A response could then be compared with both keys, with the better match used as the classification for that response. This would produce a trichotomous classification of right responses, wrong responses, and unrecognized responses needing human judgment.

Considerations for Using FRST as a Production Scoring System

One of the benefits of this study is that we can now better determine what will be required to move **FRST** into a production environment. With **FRST** as a prototype, many of the problems that were encountered could be addressed with ad hoc fixes. In the production setting, more effective solutions will be required. The problems and some potential solutions can be considered in terms of the scoring process.

(1) Extract data from the examinee record. This step requires no human intervention.

(2) Correct spelling. In terms of a large-scale implementation, spelling correction would best be handled automatically. At present we know of no spelling correctors that function in this fashion, so we would have to implement this ourselves. It is entirely possible that we could adapt an off-the-shelf product for this purpose. Such a spelling corrector would have to be studied for accuracy. It is also likely that an off-the-shelf spelling corrector would require human intervention to ensure that suggested corrections were acceptable.

A second approach is to build a spelling correction capability into the F-H test interface. This spelling correction module would alert the student to misspelled words and suggest correct spellings. At a minimum, this facility would greatly reduce the number of spelling errors encountered at the scoring stage.

(3) Filter any extraneous data, and format data for FRST. This task is currently a requirement for **FRST** but may not be a requirement in a production version. Presently all punctuation is filtered out and numbers are converted to word sequences. Then responses are processed into a form acceptable to the program. This task can be fully automated.

(4) Resolve pronouns. In this study, pronoun resolution was done manually. A more satisfactory approach would be to have a semiautomatic procedure that required minimal human involvement for verifying automatically selected resolvents. Because the F-H passages are short and the number of possible resolvents for any one referent is limited, the automatic component of this procedure should work very well.

(5) Extract the training sample. This can be done without human intervention.

(6) Train FRST. Logistically, this task involves several subtasks. One of the subtasks will be to teach test developers to train **FRST**. Presently **FRST's** interface is inadequate for general use, so it would need to be re-designed. Among other things, test developers would need to be shown how to use a rubric to create appropriate semantic classes for a set of responses and assign semantic classes to response elements. For this we would envision a one- or two-day hands-on course using **FRST**. Test developers would then assume the responsibility for training **FRST** and for performing routine quality control over the training process. Quality control would be aimed at ensuring that canonicalizations accurately capture the essence of the responses in the training sample.

(7) Execute stage-two processing to create scoring key. Human judges would be required to review, verify, and possibly modify the contents of the scoring key.

(8) Score all responses. In a production environment, scoring would require human judges for initial quality control, for ongoing updates to the scoring key, and for ongoing quality control. Initial quality control would involve verifying that the scoring

key was operating effectively. This would require running it against a new sample of responses, comparing the scores to those independently generated by human judges, adjusting the key, and then repeating the process each time with a new sample of responses until agreement between the machine and human scores reached an acceptable level. As noted, for purposes of a semiautomatic system, tuning would need to focus on accurately detecting all the truly wrong responses so that all responses classified as right could be assumed to be correct.

In operational scoring, the automatic scoring component (ASC) would operate as described above, with one change. When a response was encountered that could not be classified, the item score for an examinee would be deferred and the response passed on to the second component, the interactive scoring interface (ISI). At the ISI the human rater would decide (1) whether the response was correct and, if so, (2) how the scoring key should be changed. If the scoring key needed to be changed, this change would be reviewed by a supervising human rater at another ISI. When the change was approved, it would be returned to the ASC. The ASC would score the response, permitting an item score to be computed for the examinee, and then search the holding bin for any other responses that could be scored using this key modification. Note that this process would not result in applying different scoring rules to different examinees because it is the first instance of a given response that is routed to the judges, with all subsequent instances held until the key is updated.

To ensure the quality of this dynamic process, responses graded as correct and incorrect by the semiautomatic system should be sampled by human judges and independently scored. This level of quality control might occur continuously, as in the use of chief readers and table leaders in the Advanced Placement Program, or periodically in the form of reliability studies.

The next figure summarizes the proposed production scoring system. The design involves multiple stations using network technology to pass data back and forth between machines. Careful consideration will have to be given to the details of how this might operate.

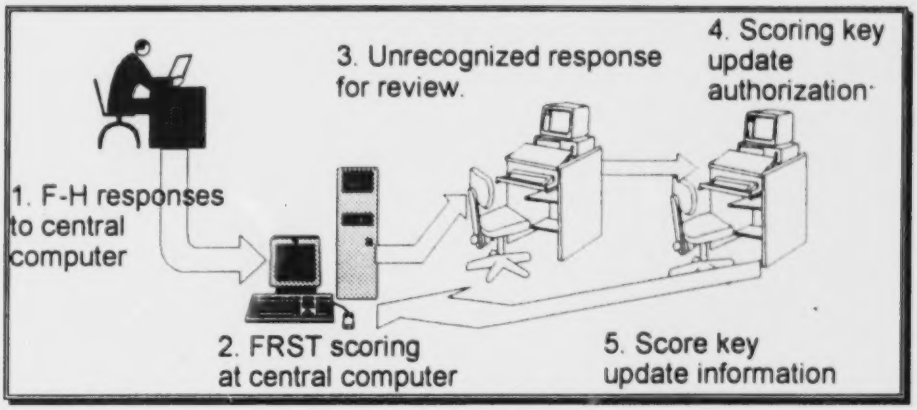
3. Unrecognized response for review.
4. Scoring key update authorization
1. F-H responses to central computer
2. FRST scoring at central computer
5. Score key update information

Figure 5 - Configuration for semiautomatic scoring

Limitations of This Study

The conditions of this study were considerably different from those that would characterize production scoring, limiting generalization of these results to that setting. Some of these differences probably worked to underestimate the accuracy of production scoring. For example, the program and tools were prototypes that should be expected to produce less positive results than would production-quality software. Also, **FRST** was used as the only scoring mechanism and not as part of a more stringent semiautomatic process. Third, the human raters' scoring rubrics were used as rough guides rather than being fully integrated into the scoring process.

Whereas some factors may have led to underestimates, the effects of other differences are more uncertain. For instance, in a real administration some students would use test-taking strategies that might produce qualitatively different responses from those encountered here. How effectively **FRST** would process those responses is not known. Also, no attempt was made to discount duplicate responses automatically. Some mechanism for dealing with these responses will be needed, if only to discourage using duplication as a test-taking trick.

Future Research

As with any experimental system, there are many steps to implementation. In this section we describe the tasks we believe are most relevant and critical.

(1) Develop spell checking mechanism. To explore the best solution, it may be necessary to implement two prototype spell checking systems. In the first, spell checking

will be done automatically as a postprocessing step. The advantage of this solution is speed. The disadvantage may be inaccuracy.

The other solution is the real-time spelling corrector that operates as an examinee is taking an exam. When the interface detects a possible spelling error, it alerts the examinee and suggests some possible alternatives. The advantage to this solution is that the data being captured should have few or no spelling errors. The disadvantage is that the spell checker might be invasive, though it could run after an examinee has finished entering the hypotheses for an item. Possible spelling errors could then be presented for remedy.

(2) Develop pronoun resolution mechanism. Here too, the question is whether a computer-assisted resolution or student-performed resolution will work better. Anaphora resolution has long been a theoretical and technical problem in natural language processing. Because the F-H passages are generally short and the number of possible referents is generally small, it may be possible to implement a semiautomatic postprocessor that requires minimal human assistance.

A second possibility is to include a pronoun resolution mechanism as part of the delivery interface. For example, if an examinee entered the word they as part of a hypothesis, the delivery interface would ask what they refers to and record the result.

(3) Create response duplication method. In the present study, we did not attempt to assign responses to conceptual classes for detecting duplication. It will be necessary to construct a postprocessing program for **FRST** that uses semantic class information to classify responses and produce scores that account for the membership of more than one hypothesis in a class. This will more closely match the scoring that human raters do for the F-H item.

(4) Develop experimental computer configuration for computer-assisted scoring. Figure 5 shows a possible configuration for a computer-assisted scoring system. The development of the prototype could help answer what the interface for semiautomatic scoring should look like and what protocol should be used to help the elements in the system communicate. The result would be a prototype system for semiautomatic scoring of F-H items using the **FRST** scoring program as the basis.

(5) Investigate variations in the scorability of the F-H item. Based on the number of iterations required to score each of the F-H items, some appear easier to process than others, in part because of differences in the linguistic variation of responses. Do some of the items score better than others because some promote use of more constrained vocabulary? Is it possible to write the items to constrain responses linguistically while not overly limiting the range of creditable ideas?

(6) Create a multiple training set version of FRST. Currently **FRST** uses a single training set containing only correct responses. It would be useful to have a version that could be trained for both correct and incorrect responses.

(7) Use a parser to augment FRST's pattern recognition capability. There is a limitation on how much can be accomplished using patterns as the primary means for scoring. **FRST**'s pattern-recognition capability can be improved with a parser, which could be used in two ways. During the training process, the parser could augment stored pattern information with syntactic information so the stored patterns could be more precisely applied. A second use is for dictionary lookup. Dictionary lookup is currently used by **FRST** only for single words because the combinatorics of phrasal lookup are prohibitive. We might limit the combinatorics by parsing the phrase, getting the parts of speech, and then using the information in selecting alternatives from the dictionary. The effectiveness of the parser will need to be explored.

Conclusions

In this study we evaluated the **FRST** program as a possible approach to scoring natural language responses to F-H items. Our evaluation has shown that **FRST** can yield scores that correlate highly with those of human raters and can recognize a large number of correct hypotheses. However, it tended to be considerably less precise in detecting truly wrong responses, classifying many of them as correct.

These results indicate that the pattern-matching approach used in **FRST** must be refined if it is to be used in a production setting. Two major ways to increase the accuracy of **FRST's** scoring are to make it a component in a semiautomatic process that employs dynamic training and to give it the ability to train on multiple classifications. The addition of a parser might also improve **FRST**'s pattern-matching capability.

References

Bejar, I. I. (1991). A methodology for scoring open-ended architectural design problems. Journal of Applied Psychology, 76, 522-532.

Bennett, R. E., & Rock, D. A. (1993). Generalizability, validity, and examinee perceptions of a computer-delivered Formulating-Hypotheses Test (RR-93-46). Princeton, NJ: Educational Testing Service.

Braun, H. I., Bennett, R. E., Frye, D, & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27, 93-108.

Carlson, S. B., & Ward, W. C. (1988). A new look at formulating hypotheses items (RR-88-12). Princeton, NJ: Educational Testing Service.

Frederiksen, N., (1959). Development of the test "Formulating Hypotheses": A progress report (Office of Naval Research Technical Report, Contract Nonr-2338(00)). Princeton, NJ: Educational Testing Service.

Frederiksen, N., & Ward, W. C. (1978). Measures for the study of creativity in scientific problem-solving. Applied Psychological Measurement, 2, 1-24.

Kaplan, R. M. (1992). Using a trainable pattern-directed computer program to score natural language item responses (RR-91-31). Princeton, NJ: Educational Testing Service.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33, 159-174.

Powers, D. E., & Enright, M. K. (1987). Analytical reasoning skills in graduate study: Perceptions of faculty in six fields. Journal of Higher Education, 58, 658-682.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. Journal of Applied Psychology, 76, 856-862.

Tucker, C. (1985). Delineation of reasoning processes important to the construct validity of the Analytical Test. Unpublished manuscript, Educational Testing Service, Princeton, NJ.

Ward, W. C., Carlson, S. B., & Woisetschlaeger, E. (1983). Ill-structured problems as multiple-choice items (RR-83-6). Princeton, NJ: Educational Testing Service.

Ward, W. C., Frederiksen, N., & Carlson, S. B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17, 11-29.

Winograd, T. (1972). Understanding natural language. New York: Academic Press.

Appendix A

Scoring Iterations and Modifications to FRST

Several modifications were made to **FRST** in the course of scoring F-H items. In addition, for some items, a repeat of the training, stage two, and scoring steps was necessary.

One of the first global changes made to **FRST** was in the way the training sets were selected. Originally, the training sets were selected by a random procedure. For the Darlington Playground and Police Officers items, we observed that **FRST**'s performance for recognizing responses was not as good as expected although it did recognize responses for which it was trained.

Key to the **FRST** process is the representativeness of the sample of responses. If the sample is not sufficiently representative, the patterns will be too limited. For this reason, the sampling process was modified to obtain a more representative training set.

The modified sampling process consisted of identifying the responses with the most unique words and using these responses for the training set.

Several other significant global modifications to **FRST** consisted of the incorporation of a morphological analyzer, the extension of the dictionary search procedure used in **FRST**, and the introduction of a capability to process general patterns.

In the initial iterations of scoring the Darlington Playground and Police Officers items, we observed that even though a word seemed to be in the lexicon, it was not classified during the pattern-creation process. Typically, this was because the word was a form of the word used in the lexicon. For example, during training, the word officer was classified as a member of the semantic class police officers. During scoring, a response possibly could have used the word officers. If this word was not explicitly encountered in the training set, it would not become part of the pattern and the response would be unrecognized. A morphological analyzer would process the word officers into a root, officer, and a suffix, s. The root is used by the pattern builder to determine if it was assigned a semantic class.

This preliminary use of the morphological analyzer, from Winograd (1972), led us to observe that it might be used at other times during the pattern-building process. The root of each entry in the lexicon is computed and stored in the lexicon during the training process. The stored root and the roots of the semantic classes are used extensively to attempt to classify a word.

The extended dictionary search occurs when a word cannot be classified by search in the lexicon. Incorporated into **FRST** is a 23,000-entry synonym dictionary. A list of

synonyms can be obtained for any word in any response. The list of words that results from lookup is then subjected to the same search procedure as the original word. If a semantic class is found for any synonym in the list, that class is used as the semantic class of the original word.

For example, suppose **FRST** encountered the hypothesis

The heat wear bulletproof vests.

The training set did contain the response (as a response pattern consisting of a series of semantic classes)

The police wear bulletproof vests.

The word police was assigned to the semantic class **law enforcement personnel**. In this example, the sense of heat is police, which was not encountered in the training set. "The heat wear bulletproof vests" is clearly a response that should be recognized. Because a lexicon search for the word heat will fail, the classification procedure resorts to calling the **synonym lookup function**. Calling on the function produces the following list of synonyms:

cop
copper
flatfoot
fuzz
heat
police officer
police
policeman

From the sentence "The police wear bulletproof vests," police has been classified. Therefore the word heat would also be classified in the semantic class **law enforcement officer**.

As a last resort in attempting to classify a response, general patterns can be used. These are generalized representations of what an acceptable response might look like. For example, in the Darlington Playground item, acceptable hypotheses might be

(1) The weather was bad outside.
(2) The weather was awful on that day.
(3) The weather was not cooperating.

The element in common in 1, 2, and 3 is the phrase "The weather was." This information could be used to construct a general pattern. That general pattern would be

(The weather was +)

The plus sign (+) is a place holder that can match any phrase in the corresponding position of a response. Any number of general patterns can be constructed and added to the scoring key. They are used only as a last resort when a pattern cannot be recognized by the normal scoring process.

Appendix B

Changes to FRST's Interface

The following modifications did not affect the training or scoring processes in **FRST,** but did come about as a result of this study. In the first study (Kaplan, 1992), the interface and its operation sufficed to score three sets of data successfully. Generally the training sets were small, and no modifications to the interface were needed . With larger training sets and a substantially greater number of semantic classes, several modifications were necessary for the interface.

In **SPAM-SCOR**, as a response was being classified, the feedback that was returned specified whether an element of a response was previously classified. An example of this is shown in the next figure, which is what is shown after an element has been assigned as semantic class.

```
(ABLE TO LIVE A LONG TIME)
 1   2  3   4 5   6
 N   N  Y   N Y   N
```

Figure B1 - Feedback given by **SPAM-SCOR** during the training process

This feedback suffices to indicate that **live** and **long** were classified but does not say anything about how they were classified. In **FRST**, this feedback was modified to include the semantic class assigned to the element. The result of this is shown below.

```
(ABLE TO LIVE       A        LONG        TIME)
 1   2  3           4        5           6
NIL  NIL life-phrase REMOVE length-phrase NIL
```

Figure B2 - Feedback given by **FRST** during the training process

In this feedback, the elements classified with **NIL** have not yet been assigned a class by the human rater. The word **life** was classified in the semantic class called **life phrase** and the word **long** was assigned to the class **length phrase**. During the training process, this information is invaluable as it immediately indicates what has been classified and how.

In addition to the modification described above, several other smaller modifications were made to the **FRST** interface as a result of scoring the F-H data. The training process took much longer for the F-H data than for the data of the first study. This was largely because of the need for a larger number of semantic classes. As training proceeds and more response elements are added to the lexicon, the time required for scoring a single response increases because the time to search the lexicon for an entry also

increases. The increased time means that an error in the training process could be disastrous. For example, if an element were incorrectly classified with no recourse for correction, it might mean that training for an item would have to be started again. Some of the training sessions required seven or eight hours, so it became important to build into the **FRST** interface a capability for recovery in the event of certain kinds of errors.

54020-11319 • Y64P 7 • 209083